Infectious Disease Modelling 7 (2022) 117–126

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Quantitative analysis of the impact of various urban socioeconomic indicators on search-engine-based estimation of COVID-19 prevalence

Ligui Wang [a, 1], Mengxuan Lin [b, 1], Jiaojiao Wang [c], Hui Chen [a], Mingjuan Yang [a], Shaofu Qiu [a], Tao Zheng [b, ***], Zhenjun Li [d, **], Hongbin Song [a, *]

[a] *Department of Infectious Disease Prevention and Control, Center for Disease Control and Prevention of Chinese People's Liberation Army, Beijing, China*

[b] *Academy of Military Medical Sciences, Academy of Military Science of Chinese PLA, Beijing, China*

[c] *The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China*

[d] *State Key Laboratory for Infectious Disease Prevention and Control, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China*

ARTICLE INFO

Article history:
Received 28 March 2022
Received in revised form 12 April 2022
Accepted 14 April 2022
Available online 20 April 2022
Handling editor: Dr Y. Shao

Keywords:
Urban socioeconomic indicators
Search engine-based estimation
Effectiveness evaluation

ABSTRACT

Numerous studies have proposed search engine-based estimation of COVID-19 prevalence during the COVID-19 pandemic; however, their estimation models do not consider the impact of various urban socioeconomic indicators (USIs). This study quantitatively analysed the impact of various USIs on search engine-based estimation of COVID-19 prevalence using 15 USIs (including total population, gross regional product (GRP), and population density) from 369 cities in China. The results suggested that 13 USIs affected either the correlation (SC-corr) or time lag (SC-lag) between search engine query volume and new COVID-19 cases (p <0.05). Total population and GRP impacted SC-corr considerably, with their correlation coefficients r for SC-corr being 0.65 and 0.59, respectively. Total population, GRP per capita, and proportion of the population with a high school diploma or higher had simultaneous positive impacts on SC-corr and SC-lag (p <0.05); these three indicators explained 37–50% of the total variation in SC-corr and SC-lag. Estimations for different urban agglomerations revealed that the goodness of fit, R^2, for search engine-based estimation was more than 0.6 only when total urban population, GRP per capita, and proportion of the population with a high school diploma or higher exceeded 11.08 million, 120,700, and 38.13%, respectively. A greater urban size indicated higher accuracy of search engine-based estimation of COVID-19 prevalence. Therefore, the accuracy and time lag for search engine-based estimation of infectious disease prevalence can be improved only when the total urban population, GRP per capita, and proportion of the population with a high school diploma or higher are greater than the aforementioned thresholds.

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: zhengtao_66@163.com (T. Zheng), lizhenjun@icdc.cn (Z. Li), hongbinsong@263.net (H. Song).
Peer review under responsibility of KeAi Communications Co., Ltd.

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.idm.2022.04.003

1. Introduction

As of 1 January 2022, the global COVID-19 pandemic had reached more than 350 million infections and 5.5 million deaths worldwide. Early estimation of COVID-19 prevalence is necessary to facilitate early response and resource allocation (Ding et al., 2021). Previously, infectious diseases were primarily monitored using reports from primary hospitals and medical institutions at a higher level (Centers for Disease C. & Prevention, 2003). This direct reporting system cannot adequately assist in monitoring the presence of emerging or unexpected infectious diseases (e.g. COVID-19) because it depends heavily on laboratory tests and has a considerable time lag.

As information on COVID-19 has primarily been acquired via the Internet during the pandemic (Hoogeveen & Hoogeveen, 2021), Internet big data-based official forecasting systems have been developed in many countries and have achieved remarkable results (Luo, 2021; Valentin et al., 2021). Search engine data, which play an integral part in Internet big data, have substantially contributed to infectious disease forecasting (Li, Liang, et al., 2020). During public health emergencies, local COVID-19 information and disease symptoms generally received the highest attention (Du et al., 2020). With increasing Internet access, people have gradually shifted from obtaining information from newspapers and television to obtaining it from the Internet. Owing to fear of the pandemic, individuals experiencing physical discomfort often enter their symptoms in a search engine to query whether their symptoms match those of COVID-19 before seeking medical advice. Therefore, the prevalence of COVID-19 can be estimated using search engine data related to COVID-19 symptoms.

Search engine-based estimation of infectious disease prevalence can be traced back to flu forecasting. A regression forecasting model was established for a flu pandemic using Google Trends (GT), and a comparison between forecast results obtained from 2003 to 2007 and the American Center for Disease Control and Prevention (CDC) data revealed that the mean correlation coefficient reached 0.90 and the flu could be forecasted one to two weeks in advance; this resulted in the formation of the Google Flu Trends (GFT) prototype (Ginsberg et al., 2009).

That study also presented two critical indicators for search engine-based estimation of infectious disease prevalence: correlation (SC-corr) and time lag (SC-lag). Correlation indicates the relationship between search volume and cases, which directly determines the estimation accuracy. Time lag indicates the time difference between the forecast and actual results, which directly determines timeliness. These two indicators are used in almost all studies related to search engine-based estimation of infectious disease prevalence.

During the COVID-19 pandemic, Kurian et al. (2020) collected the GT index for 10 COVID-19 keywords from January to April 2020 for all the states of the United States of America and analysed their correlations with COVID-19 cases. Their results revealed that the correlation coefficients between ‘Face mask’, ‘Lysol’, and ‘COVID stimulus check’ and the COVID-19 pandemic 16 days before the first COVID-19 cases were reported were 0.88, 0.92, and 0.79, respectively (Kurian et al., 2020). Further, Li, Chen, et al. (2020) retrospectively analysed the Baidu index and found that the correlation coefficients between keywords and daily incidence of COVID-19 were more than 0.89, with a time lag of 6–12 d, thereby exhibiting remarkable correlations.

Jimenez et al. (2020) summarised the studies conducted on the search engine data-based estimation of COVID-19 prevalence (as of August 2020) and found that the correlations were predominantly estimated using the GT and Baidu indices. In other studies, ten or fewer keywords were selected and their significant correlations were noted before the COVID-19 outbreak, and the minimum and maximum time lags were 1–3 d (Husnayain et al., 2020) and 18–22 d (Lu & Reis, 2020), respectively.

HealthMap is one of the most widely known global infectious disease forecasting, surveillance, and early warning systems. Operated by Boston Children's Hospital, it acquires Internet big data (Brownstein & Freifeld, 2007) from various sources (e.g. social media, news reports, online search queries) using artificial intelligence technologies to monitor disease outbreaks. The initial alarm for the global COVID-19 pandemic was signalled by HealthMap, which demonstrates its accuracy and predictive capability (Cho, 2020).

Studies on search engine-based estimation of infectious disease prevalence are subject to defects and uncontrollable factors. For example, GFT failed to estimate the H1N1 pandemic in 2009 (Olson et al., 2013) and overestimated the severity of the flu pandemic in the United States of America in 2013 (Butler, 2013). Multiple factors affect the search habits of individuals, which then impact the estimation accuracy. One such fundamental factor is that the search habits of people from different cities may not be similar and may even vary substantially. Although cities differ in size, most estimation studies integrate almost all cities to form datasets rather than analysing them separately.

As demonstrated by some phenomena during the COVID-19 pandemic, the spatial distribution characteristics of search volumes are related to the regional population and economic level. For example, the pandemic search volume is relatively high and search habits are complicated in economically developed and densely populated areas (e.g. Beijing, Shanghai, Guangdong, Sichuan, and Hubei) (Zhu et al., 2020). Therefore, temporal characteristics and regional differences should be considered when estimating COVID-19 prevalence using search engine data.

In our literature review, we found no studies that quantitatively analysed the impact of USIs[2] on the search engine-based estimation of COVID-19 prevalence. To address this issue, we quantitatively analysed the impact of various USIs on SC-corr and SC-lag based on COVID-19 epidemiological data from 369 cities in China. The COVID-19 prevalence in cities of

[2] Abbreviations: urban socioeconomic indicator (USI); gross regional product (GRP); correlation (SC-corr); time lag (SC-lag); correlation coefficients r; goodness of fit R^2.

different sizes was estimated using the search engine and a direct comparison of estimation accuracies and threshold calculations was performed. Our study will provide a reference for other studies on search engine-based estimation of infectious disease prevalence.

2. Methodology

2.1. Data collection and screening

We collected data from 369 cities in 31 provinces and regions of China. They consisted of 15 USIs: total population; population density (total urban population divided by urban area); GRP (GDP of a city in one year); proportions of primary, secondary, and tertiary sectors in GRP (primary sector: agriculture; secondary sector: handicraft; tertiary sector: modern service or business); GRP per capita; public budget revenue (tax-based fiscal revenue); public budget expenditure (fiscal expenditure); education expenditure; science and technology expenditure; rate of natural increase (natural increase in urban population in one year divided by average urban population over the same period); proportion of the population with a high school diploma or higher; urbanisation rate (urban population divided by total urban population); and proportion of the population aged 0–39 y.

We obtained the Baidu search index for the 369 cities during the COVID-19 pandemic using search engine web crawlers. Based on the topics (e.g. symptoms, pathogens, geographical areas, and actions against COVID-19), 32 COVID-19-related keywords were selected (see Supplementary Materials Table S1).

We obtained information regarding daily new COVID-19 cases in the 369 cities from the websites of provincial health commissions in China (see Supplementary Materials Table S2). The start date was January 20, 2020, as statistics on new COVID-19 cases before this date were not recorded in most cities. Since March 2020, the number of daily new COVID-19 cases in most cities in China was zero for several consecutive days. As data with no new COVID-19 cases are insignificant to the results, only new COVID-19 cases occurring in the early stage of the COVID-19 outbreak were selected. We selected 44 cities of different sizes as USI samples where more than 100 cumulative COVID-19 cases were reported; the daily new COVID-19 cases and Baidu index for keywords in each city from January 20, 2020 to February 29, 2020 were selected as the samples for the search engine-based estimation.

2.2. Correlation and time lag

Correlation denotes the relationship degree (or connectedness) between two variables, expressed quantitatively by the correlation coefficient. Time lag, which denotes the interval between two variables, is expressed quantitatively by lag time. In this study, the correlation between the Baidu index for search keywords and new COVID-19 cases in each city (hereinafter SC-corr) and the time lag between the Baidu index for search engine keywords and new COVID-19 cases (hereinafter SC-lag) occurred sequentially. The following rules were followed: correlation coefficient (r) values of 0.6 or more, 0.4–0.6, and less than 0.4 indicated high, medium, and low correlation, respectively. Time lags of 6 or more, 4–6, and less than 4 d indicated high, medium, and low lag, respectively.

2.3. Fitted model

This study quantitatively analysed the impact of various USIs on the estimation of COVID-19 prevalence based on search engine data using linear fitting. As USIs involve multiple dimensions and substantial collinearity is present between data of different dimensions, we investigated the impacts using univariate linear fitting on the data. The fitting function is as follows:

$$y_i = \alpha_i + \beta_i x_i + \varepsilon_i \tag{1}$$

where i, the number of variables, and y_i, the output variable, denote the fitted values of Pearson's correlation coefficient and lag time for daily new COVID-19 cases and the Baidu index for keywords for 44 cities, respectively. α_i denotes a constant term, β_i denotes a fitting coefficient, x_i denotes a USI, and ε_i denotes an error.

When the correlation coefficient and lag time were considered simultaneously, we performed the analysis using three-dimensional surface fitting. The fitting function is as follows:

$$z_i = \alpha_i + \mu_i x_i + \vartheta_i y_i + \sigma_i x_i y_i + \mu_i {x_i}^2 + \tau_i {y_i}^2 + \varepsilon_i \tag{2}$$

where i denotes the number of variables; z_i denotes the output variable; α_i denotes a constant term; μ_i, ϑ_i, σ_i, μ_i, and τ_i denote a fitting coefficient; x_i and y_i denote USI; and ε_i denotes an error.

2.4. Urban classification criteria

During estimation, we classified cities of different sizes into urban agglomerations and used them as the datasets to simulate the actual estimates. We defined the urban classification indicators as follows:

$$C = \prod_{i=1}^{b}\left(1 + \frac{x_i - x_{i\ Min}}{x_{i\ Max} - x_{i\ Min}}\right) \tag{3}$$

where C denotes the size indicator for our classification and b denotes the number of urban indicators that can have a substantial and simultaneous effect on correlation and time lag. During estimation, USIs that could affect SC-corr and SC-lag significantly and simultaneously were substituted into Eq. (3) to calculate the urban classification indicator C. Based on the values of C, the cities were divided into five levels: Level V (super cities), Level IV (megacities), Level III (big cities), Level II (medium-sized cities), and Level I (small cities) (see Supplementary Materials Table S3).

2.5. Estimation model

We analysed the quantitative relationship between the search engine data and the number of new COVID-19 cases and established estimation models for all cities. People often use multiple related keywords rather than one keyword when searching for information about COVID-19; as we used COVID-19-related keyword sets, this resulted in data multicollinearity. Consequently, the search index for each keyword was highly correlated with each other (see Supplementary Materials Fig. S1). We eliminated collinearity and generated sparse solutions using L1 regularisation (Lasso regression) to ensure accuracy, robustness, and stability of results. The loss function for the Lasso regression model used in this study is as follows:

$$J = \frac{1}{2m}\sum_{t=1}^{m}\left(\log(Y_t) - \sum_{i=1}^{n}\beta_{s,i}\lg(X_{t+s,i}) - \mu\right)^2 + \lambda\left(\sum_{i=1}^{n}\beta_{s,i}\right) \tag{4}$$

where λ denotes the penalty coefficient, $\beta_{s,i}$ and $X_{t+s,i}$ respectively denote the regression coefficient and relative index for keyword i after a lag of s day(s), n denotes the number of selected keywords, and m denotes the sample size.

3. Results

3.1. Analysis results of SC-corr and SC-lag

Based on the comparisons between the Baidu index for 32 keywords and correlation coefficients for daily new COVID-19 cases in China (see Supplementary Materials Table S1), a trend analysis was performed on the Baidu index for five keywords with $r > 0.8$ and daily new COVID-19 cases (Fig. 1).

Fig. 1 intuitively shows that the wave shapes for the Baidu index for the five keywords—namely, 'Fever', 'Cough', 'Fatigue', 'Coronavirus', and 'Novel coronavirus'—and daily new COVID-19 cases in China remained similar. In addition, the wave peak position indicates that a lag time was present. The presence of the lag time is attributable to the fact that the incubation period for COVID-19 is 1–14 d and the Baidu index is likely to rise a few days before a surge in the number of new COVID-19 cases. We obtained the mean correlation coefficient set with a lag time of 1–14 d by moving the time series for the Baidu index for five keywords and daily new COVID-19 cases in each city and selecting the maximum correlation coefficient and corresponding time as the correlation coefficient and the lag time for each city, respectively. Thereafter, we calculated and screened out the correlation coefficients and lag times of the 44 cities used in this study (see Supplementary Materials Table S4).

From the results, the proportions of cities with a high, medium, and low correlation were 34.1%, 36.4%, and 29.5%, respectively, and the proportions of cities with a high, medium, and low lag were 22.7%, 38.6%, and 38.7%, respectively in the 44 cities of different sizes where more than 100 cumulative confirmed COVID-19 cases were reported in China (Fig. 2).

It is noteworthy that, of the 44 cities, the Pearson correlation coefficient values and lag times were surprisingly low solely for Jining. This is because most COVID-19 infections occurred in a Jining prison on February 22, 2020 and led to more than 200

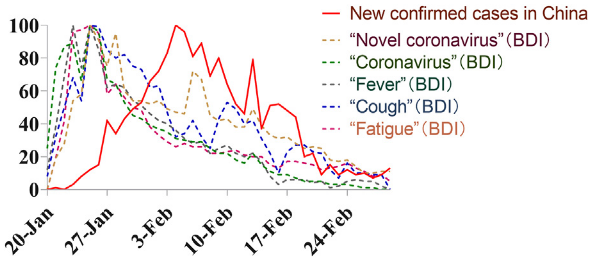

Fig. 1. Baidu index for five keywords with maximum correlation coefficient and new COVID-19 cases in in 2020. Because the Baidu index and confirmed cases vary extensively in terms of unit and order of magnitude, the Baidu index and new confirmed cases were standardised to a scale ranging from 0 to 100 for an intuitive comparison.

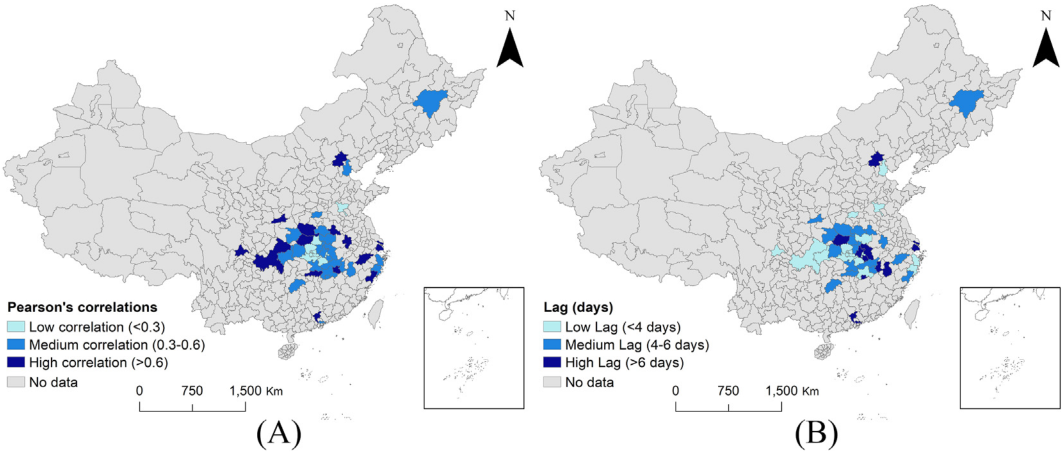

Fig. 2. Correlation and time lag between the Baidu index and new cases in cities heavily affected by COVID-19 in China. A) SC-corr distribution. B) SC-lag distribution.

new COVID-19 cases on the same day. No more than 10 daily new COVID-19 cases were reported within the selected date range, excluding February 22, 2020, for which a surge in the number of new COVID-19 cases was reported. This abnormal data fluctuation led to confusion about the correlation coefficient and lag time. Therefore, to avoid the impact on the results, Jining was considered a bad datapoint and rejected; thus, the impact of USIs was analysed using data from the remaining 43 cities only.

3.2. Analysis results of the impacts of USIs on SC-corr and SC-lag

We conducted the fitting analysis using 15 USIs and compared them with their SC-corr and SC-lag, respectively (Fig. 3). During the estimation of COVID-19 prevalence, a higher SC-corr and SC-lag indicated a more accurate estimation result and greater practical significance. During the fitting analysis, we selected the aforementioned strong correlation (r >0.6) and strong time lag (lag time of >6 d) as the thresholds for determining good estimation results and took the first city that exhibited a greater correlation and time lag than the thresholds as the threshold city (Table 1).

Regarding correlation, the fitting equations for the total population, population density, GRP, proportion of tertiary sector in GRP, GRP per capita, public budget revenue, public budget expenditure, education expenditure, science and technology expenditure, proportion of population with a high school diploma or higher, and urbanisation rate passed the significance test, proving that these USIs substantially affect SC-corr. The correlation coefficients for total population and GRP, r, were 0.65 and 0.59, respectively. Only the goodness of fit R^2 values for total population, GRP, public budget revenue, and education expenditure exceeded 0.5, and these four USIs respectively explain 41.95%, 34.84%, 22.79%, and 23.20% of the data fluctuations, suggesting that they had a greater impact than the other USIs.

The results for the impact of various USIs on time lag are very different. The fitting equations for population density, GRP, GRP per capita, proportion of high school and above, urbanisation rate, and proportion of population aged 0–39 years passed the significance test, proving that these USIs substantially affect SC-lag. However, for most USIs, with r values being less than 0.45, only the proportion of population with a high school diploma or higher had a goodness of fit of more than 0.2. These results indicate that USIs have a greater impact on SC-corr than SC-lag.

The accuracy of estimation of COVID-19 prevalence was subject to SC-corr and SC-lag. After considering the correlation coefficient and lag time simultaneously, it was found that the USI results converged significantly, only three of the 15 USIs (total population, GRP per capita, and population with a high school diploma or higher) passed the significance test. We listed the threshold cities for SC-corr and SC-lag in Table 1 and used them as the reference standards for screening data during COVID-19 forecasting. If the USI values were below the thresholds, SC-corr and SC-lag for that city would be considered weak, and estimating COVID-19 prevalence in those cities using USIs as the data source would not have distinct significance, and could even affect the objectivity of the results.

3.3. Impact of USIs on estimation

Within the selected date range, Lasso regression models were respectively established for cities of five levels using the Baidu index for the five search keywords ‘Fever’, ‘Cough’, ‘Fatigue’, ‘Coronavirus’, and ‘Novel coronavirus’, and daily new COVID-19 cases (Fig. 4). Thereafter, the estimation performances for cities of different levels were compared directly within the selected date range (Table 2).

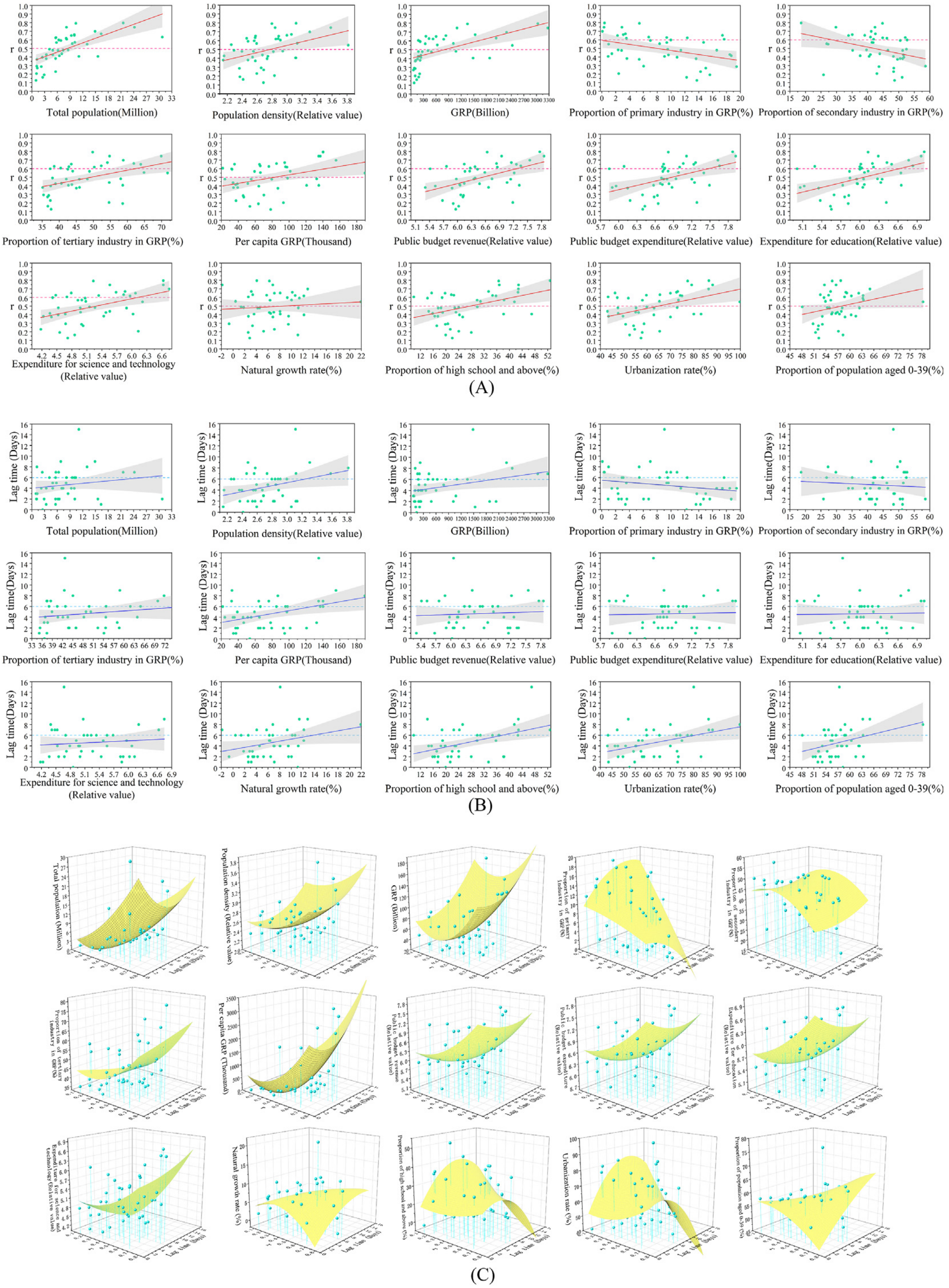

Fig. 3. Impact of USIs on correlation (SC-corr) and time lag (SC-lag). A) Impact of USIs on SC-corr. Solid red line indicates linear fitting results, and dashed red line indicates the dividing line between high correlation and medium correlation. B) Impact of USIs on SC-lag. Solid blue line indicates linear fitting results, and dashed blue line indicates the dividing line between high hysteresis and medium hysteresis. C) Impact of USIs on SC-corr and SC-lag. Yellow surface indicates fitting results and grey area indicates the 95% confidence interval belt.

The results demonstrate that it is feasible to conduct search engine-based estimation of COVID-19 prevalence for cities of different sizes ($p < 0.001$); however, the estimation accuracies varied substantially. For USIs from Level V cities, the search

Table 1
Fitted model for USIs and SC-corr as well as SC-lag.

Indicators	Total population (million)	Population density (relative value)	GRP (billion)	Proportion of primary sector in GRP (%)	Proportion of secondary sector in GRP (%)	Proportion of tertiary sector in GRP (%)	GRP per capita (thousand)	Public budget revenue (relative value)	Public budget expenditure (relative value)	Education expenditure (relative value)	Science and technology expenditure (relative value)	Rate of natural increase (%)	Proportion of population with a high school diploma or higher (%)	Urbanisation rate (%)	Proportion of population aged 0–39 years (%)
Strong correlation (critical value: $r = 0.6$)															
Threshold	14.01	3.271	1593	NA	NA	62.56	143.5	7.28	7.48	6.63	6.13	NA	40.87	82.70	NA
Threshold city	Guangzhou (14.90)	Guangzhou (3.302)	Tianjin (1881)	NA	NA	Hangzhou (63.90)	Guangzhou (155.4)	Chongqing (7.36)	Shenzhen (7.63)	Guangzhou (6.64)	Guangzhou (6.21)	NA	Shenzhen (41.55)	Tianjin (83.15)	NA
r	0.65	0.41	0.59	−0.27	−0.36	0.49	0.39	0.50	0.47	0.50	0.47	0.09	0.45	0.45	0.29
Fitting R^2	0.4195	0.1663	0.3484	0.1435	0.1070	0.2209	0.1490	0.2279	0.2028	0.2320	0.2049	0.0082	0.1997	0.2018	0.0835
Fitted p value	<0.0001	<0.05	<0.0001	Not significant	Not significant	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	Not significant	<0.05	<0.05	Not significant
Strong hysteresis (critical value: *lag time* = 6)															
Threshold	NA	3.239	1984	NA	NA	NA	127.7	NA	NA	NA	NA	NA	37.71	79.9	64.74
Threshold city	NA	Guangzhou (3.302)	Chongqing (2036)	NA	NA	NA	Ningbo (132.6)	NA	NA	NA	NA	NA	Tianjin (38.13)	Wuhan (80.04)	Guangzhou (65.01)
r	0.17	0.33	0.32	−0.21	−0.08	0.17	0.39	0.07	0.03	0.02	0.10	0.30	0.45	0.40	0.32
Fitting R^2	0.0277	0.1103	0.1020	0.0229	0.0178	0.0073	0.1556	0.0097	0.0042	0.0007	0.0121	0.0873	0.2018	0.1563	0.1044
Fitted p value	Not significant	<0.05	<0.05	Not significant	Not significant	Not significant	<0.05	Not significant	Not significant	Not significant	Not significant	Not significant	<0.05	<0.05	<0.05
Strong correlation and strong hysteresis															
Threshold	Guangzhou (14.90)	NA	NA	NA	NA	NA	Guangzhou (155.4)	NA	NA	NA	NA	NA	Xi'an (42.67)	NA	NA
Fitting R^2	0.50	0.17	0.26	0.09	0.08	0.21	0.45	0.22	0.20	0.23	0.19	0.03	0.37	0.27	0.09
Fitted p value	<0.05	Not significant	Not significant	Not significant	Not significant	Not significant	<0.05	Not significant	Not significant	Not significant	Not significant	Not significant	<0.05	Not significant	Not significant

engine-based estimation of the COVID-19 prevalence model performed exceptionally well (adjusted R^2 value = 0.711). For USIs from Level IV cities, the adjusted R^2 value fell to 0.601. For USIs from Levels I–III cities, the estimation accuracies were low (adjusted R^2 values < 0.55), resulting in unsatisfactory estimation results. These results differed because the search habits and patterns of citizens in cities of different sizes vary substantially, and the search engine data from cities of greater sizes are more closely related to their local COVID-19 prevalence. Typically, greater values of USIs indicate a higher estimation accuracy and a lower error.

The results show that low-level cities affect the overall result accuracy of estimation if all cities are combined into a dataset without urban size differentiation. If the adjusted R^2 value estimated to be > 0.6 met the acceptable minimum requirement, the low-level cities may be necessarily rejected, and only USIs from Level IV cities would be included in the dataset to ensure estimation accuracy. Therefore, we concluded that total population, GRP per capita, and proportion of the population with a high school diploma or higher can be used as variables for estimation of COVID-19 prevalence only when they are more than 11.08 million, 120,700, and 38.13%, respectively.

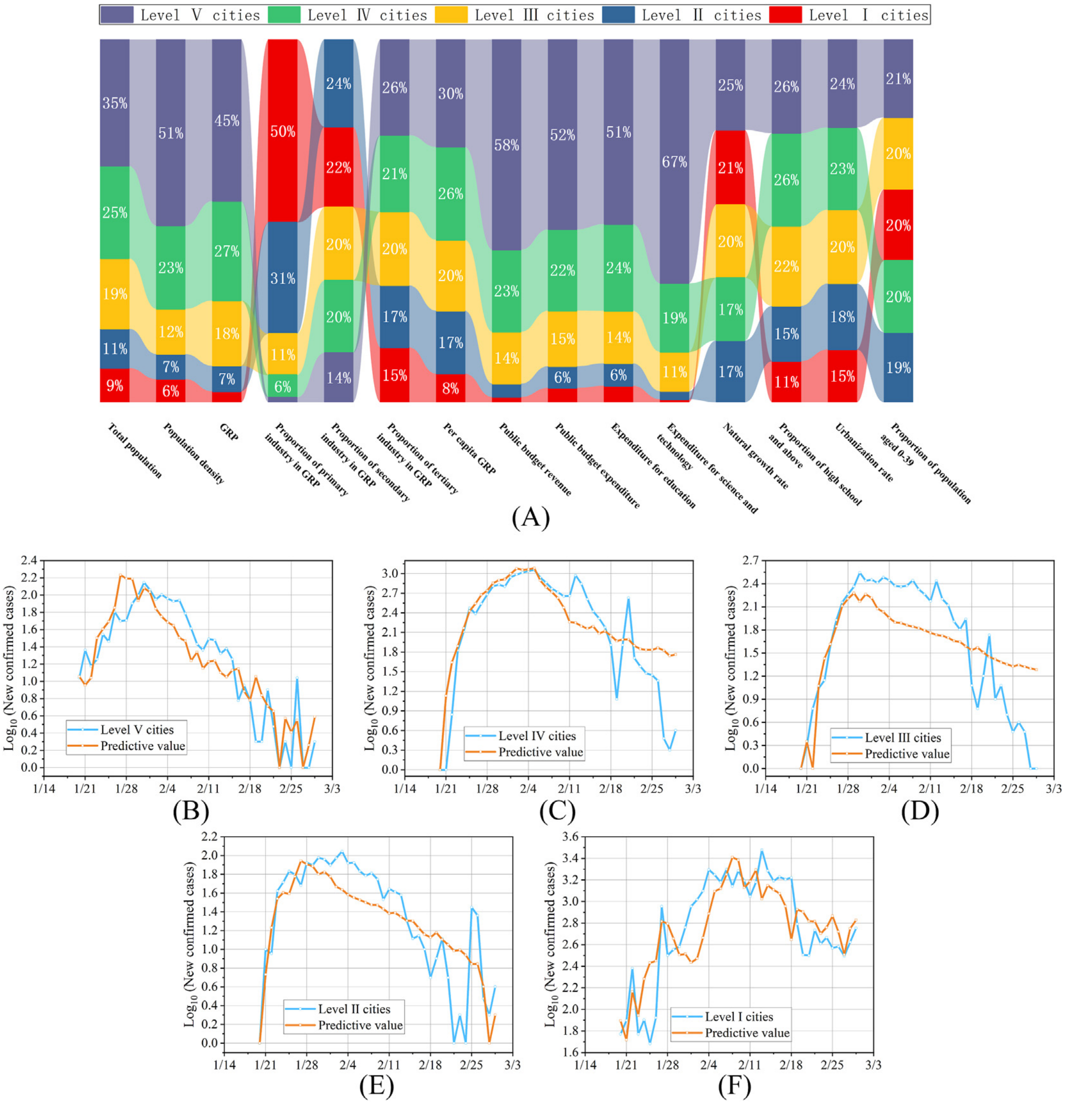

Fig. 4. Effects of search engine data-based actual estimation of COVID-19 prevalence in cities of different levels. A) Proportions of USIs from cities of different levels. The respective medians of 15 USIs from Levels I–V cities were selected to avoid the impact of abnormal values and ensure objective comparisons. B) Estimation results for Level V cities. C) Estimation results for Level IV cities. D) Estimation results for Level III cities. E) Estimation results for Level II cities. F) Estimation results for Level I cities.

Table 2
Size range of urban agglomerations of different levels and model performance during the actual estimation.

Model performance	Level I cities	Level II cities	Level III cities	Level IV cities	Level V cities
Classification criteria	$C < 1.9$	$1.9 < C < 2.8$	$2.8 < C < 3.7$	$3.7 < C < 4.6$	$C > 4.6$
Total population (million)	0.97–10.01	1.07–9.52	8.15–30.75	11.08–15.57	14.90–24.18
GRP per capita (thousand)	21.6–79.5	59.2–132.6	65.9–140.2	120.7–135.1	135.0–190.0
Proportion of population with a high school diploma or higher (%)	9.90–27.98	19.74–34.67	21.70–42.67	38.13–46.97	41.55–52.72
Lasso regression *k* value	0.28	0.36	0.27	0.16	0.1
p value	<0.001	<0.001	<0.001	<0.001	<0.001
Adjusted R^2 value	0.460	0.493	0.508	0.601	0.711
RMSE	0.575	0.527	0.346	0.290	0.305

4. Discussion

This study pioneered the analysis of the impact of USIs on search engine-based estimation of COVID-19 prevalence. With growing access to the Internet and the rise of big data, the estimation of infectious disease prevalence based on search engine data, characterised by high speed, accurate results, and universal application scope, can be performed extensively. For example, the utilisation of such estimation in combination with traditional disease surveillance systems during the COVID-19 pandemic demonstrated its substantial benefits and practical significance, resulting in numerous studies on search engine-based estimation. However, almost all studies took search engine data from a country or region as a whole, and did not consider the remarkable differences in search engine data between cities of different sizes, thereby resulting in low estimation accuracy.

We used 15 USIs from 369 cities of different sizes in 31 provinces in China, of which 44 cities in 15 provinces were heavily affected by COVID-19. Specifically, 15 USIs, which represented a large sample set, were used for each city to produce universally applicable results. We quantitatively analysed and calculated SC-corr and SC-lag for cities of different sizes, and performed the fitting analyses of USI and SC-corr as well as USI and SC-lag; USIs that affected SC-corr and SC-lag were screened, and relevant threshold cities were determined by calculation. The study results can help public health authorities select varying strategies for cities of different sizes in terms of forecasting and early warning of COVID-19 to effectively improve estimation accuracy.

The study findings revealed that search engine data from larger cities are more closely related to COVID-19 prevalence and search engine-based estimation is feasible. The estimation can be conducted more than six days in advance only for a few cities. The total population, GRP, public budget revenue, and education expenditure significantly affected the estimation results; total population determines the urban size, GRP determines the degree of urban economic development, public budget revenue determines the income level of urban residents, and education expenditure determines the number of urban schools and education quality.

We believe that a reasonable factor involving high-level government administration, improved policies, a sophisticated query platform for COVID-19 information, high quality of life of residents, and their growing willingness to enquire about COVID-19 information on the Internet to determine whether they should go to a hospital for further examinations in cities with high USI values led to the considerable correlation between search engine data and local COVID-19 prevalence.

We conducted the estimations using search engine data and the number of new cases in urban agglomerations of different levels, respectively. Our results verified the analysis results of SC-corr and SC-lag, and revealed that greater urban size correlates with higher estimation accuracy, which is consistent with the conclusion. In addition, we presented thresholds for USIs that may be referenced for screening datasets during estimation; cities with the total population, GRP per capita, and proportion of the population with a high school diploma or higher exceeding 11.08 million, 120,700, and 38.13%, respectively, exhibited a higher goodness of fit R^2 in COVID-19 prevalence estimation.

5. Conclusion

This study is the first to quantitatively analyse the impact of 15 USIs on search engine-based estimation of COVID-19 prevalence. The results reveal that cities of different sizes exhibit varying degrees of accuracy and time lags for search engine-based estimation of COVID-19 prevalence, and the accuracy of search engine-based estimation of infectious disease prevalence could meet the expectation only when the total urban population, GRP per capita, and proportion of population with a high school diploma or higher exceeded 11.08 million, 120,700, and 38.13%, respectively. The accuracy and time lag for search engine-based estimation of infectious disease prevalence were effectively improved based on our study. This study provides a reference for the practical implementation of search engine-based estimation of infectious disease prevalence.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key R&D Program of China (2021YFC2302004) and the National Natural Science Foundation of China (Grant Nos. 72074209, 72042018, 71621002).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.idm.2022.04.003.

References

Brownstein, J. S., & Freifeld, C. C. (2007). HealthMap: The development of automated real-time internet surveillance for epidemic intelligence. *Euro Surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 12*(11). E071129.071125-E071129.071125. <Go to ISI>://MEDLINE:18053570.

Butler, D. (2013). When Google got flu wrong: US outbreak foxes a leading web-based method for tracking seasonal flu. *Nature, 494*(7436), 155–157.

Centers for Disease, C., & Prevention. (2003). Update: Severe acute respiratory syndrome–United States, May 21, 2003. *MMWR. Morbidity and mortality weekly report, 52*(20), 466–468. Go to ISI>://MEDLINE:12807079.

Cho, A. (2020). COVID-19 AI systems aim to sniff out coronavirus outbreaks. *Science, 368*(6493), 810–811. https://doi.org/10.1126/science.368.6493.810

Ding, W. P., Nayak, J., Swapnarekha, H., Abraham, A., Naik, B., & Pelusi, D. (2021). Fusion of intelligent learning for COVID-19: A state-of-the-art review and analysis on real medical data. *Neurocomputing, 457*, 40–66. https://doi.org/10.1016/j.neucom.2021.06.024

Du, H. F., Yang, J., King, R. B., Yang, L., & Chi, P. L. (2020). COVID-19 Increases online searches for emotional and health-related terms. *Applied Psychology-Health and Well Being, 12*(4), 1039–1053. https://doi.org/10.1111/aphw.12237

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–U1014. https://doi.org/10.1038/nature07634

Hoogeveen, M. J., & Hoogeveen, E. K. (2021). Comparable seasonal pattern for COVID-19 and flu-like illnesses. *One Health, 13*. https://doi.org/10.1016/j.onehlt.2021.100277. Article 100277.

Husnayain, A., Fuad, A., & Su, E. C.-Y. (2020). Applications of Google search trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in taiwan. *International Journal of Infectious Diseases, 95*, 221–223. https://doi.org/10.1016/j.ijid.2020.03.021

Jimenez, A. J., Estevez-Reboredo, R. M., Santed, M. A., & Ramos, V. (2020). COVID-19 symptom-related Google searches and local COVID-19 incidence in Spain: Correlational study. *Journal of Medical Internet Research, 22*(12). https://doi.org/10.2196/23518. Article e23518.

Kurian, S. J., Bhatti, A. U. R., Alvi, M. A., Ting, H. H., Storlie, C., Wilson, P. M., Shah, N. D., Liu, H., & Bydon, M. (2020). Correlations between COVID-19 cases and Google trends data in the United States: A state-by-state analysis. *Mayo Clinic Proceedings, 95*(11), 2370–2381. https://doi.org/10.1016/j.mayocp.2020.08.022

Li, C., Chen, L. J., Chen, X., Zhang, M., Pang, C. P., & Chen, H. (2020). Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Euro Surveillance, 25*(10), 7–11. https://doi.org/10.2807/1560-7917.Es.2020.25.10.2000199. Article 2000199.

Li, K., Liang, Y., Li, J., Liu, M., Feng, Y., & Shao, Y. (2020). Internet search data could be used as novel indicator for assessing COVID-19 epidemic. *Infectious Disease Modelling, 5*, 848–854. https://doi.org/10.1016/j.idm.2020.10.001

Luo, J. (2021). Forecasting COVID-19 pandemic: Unknown unknowns and predictive monitoring. *Technological Forecasting and Social Change, 166*. https://doi.org/10.1016/j.techfore.2021.120602. Article 120602.

Lu, T., & Reis, B. Y. (2020). Internet search patterns reveal clinical course of COVID-19 disease progression and pandemic spread across 32 countries. *medRxiv* https://doi.org/10.1101/2020.05.01.20087858.

Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., & Simonsen, L. (2013). Reassessing Google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Computational Biology, 9*(10), Article e1003256.

Valentin, S., Mercier, A., Lancelot, R., Roche, M., & Arsevska, E. (2021). Monitoring online media reports for early detection of unknown diseases: Insight from a retrospective study of COVID-19 emergence. *Transboundary and Emerging Diseases, 68*(3), 981–986. https://doi.org/10.1111/tbed.13738

Zhu, B. R., Zheng, X. Q., Liu, H. Y., Li, J. Y., & Wang, P. P. (2020). Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. *Chaos, Solitons & Fractals, 140*. https://doi.org/10.1016/j.chaos.2020.110123. Article 110123.